

Sampling Requirements and Accelerated Schemes for Sparse Linear Regression with Orthogonal Least-Squares

Abolfazl Hashemi, *Student Member, IEEE*, and Haris Vikalo, *Senior Member, IEEE*

Abstract—The Orthogonal Least Squares (OLS) algorithm sequentially selects columns of the coefficient matrix to greedily find an approximate sparse solution to an underdetermined system of linear equations. Previous work on the analysis of OLS has been limited; in particular, there exist no guarantees on the performance of OLS for sparse linear regression from random measurements. In this paper, the problem of inferring a sparse vector from random linear combinations of its components using OLS is studied. For the noiseless scenario, it is shown that when the entries of a coefficient matrix are samples from a Gaussian or a Bernoulli distribution, OLS with high probability recovers a k -sparse m -dimensional sparse vector using $\mathcal{O}(k \log m)$ measurements. Similar result is established for the bounded-noise scenario where an additional condition on the smallest nonzero element of the unknown vector is required. Moreover, generalizations that reduce computational complexity of OLS and thus extend its practical feasibility are proposed. The generalized OLS algorithm is empirically shown to outperform broadly used existing algorithms in terms of accuracy, running time, or both.

Index Terms—linear regression, compressed sensing, sparse reconstruction, greedy algorithm, orthogonal least-squares

I. INTRODUCTION

THE task of estimating a sparse vector from a few linear combinations of its components, readily cast as the problem of finding a sparse solution to an underdetermined system of linear equations, is encountered in many practical scenarios, including sparse linear regression [1], compressed sensing [2], sparse channel estimation in communication systems [3], [4], compressive DNA microarrays [5] and a number of other applications in signal processing and machine learning [6]–[11]. Consider the linear measurement model

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\nu}, \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^n$ denotes the vector of observations, $\mathbf{A} \in \mathbb{R}^{n \times m}$ is the coefficient matrix (i.e., a collection of features) assumed to be full rank (generally, $n < m$), $\boldsymbol{\nu} \in \mathbb{R}^n$ is the additive observation noise vector, and $\mathbf{x} \in \mathbb{R}^m$ is an unknown vector assumed to have at most k non-zero components (i.e., k is the sparsity level of \mathbf{x}). Finding a sparse approximation to \mathbf{x} leads to a cardinality-constrained least-squares problem

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \quad \text{subject to} \quad \|\mathbf{x}\|_0 \leq k, \quad (2)$$

known to be NP-hard; here $\|\cdot\|_0$ denotes the l_0 -norm, i.e., the number of non-zero components of its argument. The high cost

of finding the exact solution to (2) motivated development of a number of heuristic methods that trade accuracy for computational efficiency. These heuristics can generally be grouped in two categories. The first set of methods facilitate computationally efficient search for a sparse solution by replacing the non-convex cardinality-constrained optimization by a sparsity-promoting l_1 -norm optimization. It was shown in [12] that such a formulation provides exact recovery of a sufficiently sparse signal from noise-free measurements under certain conditions on the problem parameters. However, while the convexity of l_1 -norm enables algorithmically straightforward sparse vector recovery by means of, e.g., iterative shrinkage-thresholding [13] or alternating direction method of multipliers [14], the complexity of such methods is often prohibitive in practice. The second set of methods for sparse approximation consist of a number of iterative and greedy heuristics that attempt to satisfy the cardinality constraint directly by successively identifying columns of the coefficient matrix which correspond to non-zero components of the unknown vector.

Among the greedy methods for sparse vector reconstruction, the orthogonal matching pursuit (OMP) algorithm [15] has attracted particular attention in recent years. Intuitively appealing due to its simple geometric interpretation, OMP is characterized by high speed and competitive performance. In each iteration, OMP selects a column of the coefficient matrix \mathbf{A} having the maximum correlation with a so-called residual vector and adds it to the set of active columns; then the projection of the observation vector \mathbf{y} onto the space spanned by the columns in the active set is used to form a residual vector needed for the next iteration of the algorithm. Numerous modifications of OMP with enhanced performance have been proposed in literature. For instance, instead of choosing a single column in each iteration of OMP, [16] selects and explores all columns having correlation with a residual vector that is greater than a pre-determined threshold. [17] employs the same idea, but instead of thresholding, a fixed number of columns are selected per iteration. [18] identifies columns with largest proximity to the residual vector, use them to find a least-squares approximation of the unknown signal, and retains only the largest entries in the resulting approximation. [19] applies a similar approach where at first a set of k columns having the largest correlation with the residual vector is identified, and then in each iteration a number of columns is added to or eliminated from the set based on their correlation with the residual vector. In [20], columns are chosen based on their mutual correlation and aggregated energy. Necessary and

sufficient conditions for exact reconstruction of sparse signals using OMP have been established. Examples of these results include analysis under Restricted Isometry Property (RIP) [21]–[23], recovery conditions based on Mutual Incoherence Property (MIP) and Exact Recovery Condition (ERC) [24]–[26], and the conditions based on the so-called submodularity ratio [27]. For the case of random measurements, performance of OMP was analyzed in [28]–[30]. Tropp et al. in [28] showed that in the noise-free scenario, $\mathcal{O}(k \log m)$ measurements is adequate to recover k -sparse m -dimensional signals with high probability. In [29], this result was extended to the case of noisy measurements in the high SNR regime under the assumption that the entries of \mathbf{A} are i.i.d Gaussian and that the length of the unknown vector approaches infinity (i.e., that work provides asymptotic analysis of the OMP performance).

The Orthogonal Least-Squares (OLS) algorithm was first introduced in the statistics literature as a forward regression scheme with applications to subset selection [31], [32]. Chen et al. [33] proposed OLS as a method for estimating parameters of generally multivariate non-linear systems which are linear in the parameters. OLS has drawn attention in recent years and its performance was analyzed in limited settings. In [34], OLS was analyzed under the Exact Recovery Condition (ERC), first introduced in [24]. Herzet et al. [35] provided coherence-based conditions for sparse recovery of signals via OLS when nonzero components obey certain decay conditions. In [36], sufficient conditions for exact recovery are stated when a subset of optimal indices is available. However, all the existing analysis and performance guarantees for OLS pertain to non-random measurements.

A. Contribution

In this paper, we establish conditions for exact recovery of the sparse vector \mathbf{x} from measurements \mathbf{y} in 1 using OLS, where the entries of the coefficient matrix \mathbf{A} are drawn at random from a Gaussian or a Bernoulli distribution. Specifically, we first present conditions which ensure that, in the noise-free scenario, OLS with high probability recovers the support of \mathbf{x} in at most k iterations. Following the framework in [28], we further find a lower bound on the probability of performing exact sparse recovery in at most k iterations and demonstrate that with $\mathcal{O}(k \log m)$ measurements OLS succeeds with probability arbitrarily close to one. Moreover, we extend the analysis to the case of noisy measurements and show that similar guarantees hold if the nonzero element of \mathbf{x} with the smallest magnitude satisfies certain condition. This condition implies that to ensure exact support recovery via OLS in the presence of additive white Gaussian noise, SNR should scale linearly with the sparsity level.

In addition to performance analysis, we propose an efficient implementation of the OLS algorithm which recursively updates the orthogonal projection operator to the span of previously selected columns and therefore the residual vector needed for the subsequent iterations. This modification of OLS is motivated by observing a recursive relation between the components of the optimal solution to the original l_0 -constrained least-squares problem 2. The resulting algorithm,

referred to as Accelerated Orthogonal Least-Squares (AOLS), is shown to be computationally superior compared to classical OLS over a wide range of problem parameters. Finally, we propose a generalization of OLS, the Generalized Orthogonal Least-Squares (GOLS), which exploits the observation that columns having strong correlation with the current residual are also likely to have strong correlation with residuals in subsequent iterations; this justifies selection of multiple columns in each iteration and formulation of an overdetermined system of linear equation having solution that is generally more accurate than the one found by OLS. Our extensive empirical studies show that GOLS is more accurate than both OLS and OMP, while being computationally feasible (in particular, faster than OLS though slower than OMP).

B. Organization

The remainder of the paper is organized as follows. In Section II, we specify the notation and overview the classical OLS algorithm. Section III presents performance guarantees for sparse linear regression from random measurements using OLS. In Section IV, we describe computationally efficient variants of OLS. Section V presents experiments which empirically verify results on sampling requirements for OLS and benchmark performance of the proposed algorithms. Finally, concluding remarks are provided in Section VI.

II. PRELIMINARIES

A. Notation

We here briefly summarize notation used in the paper. Bold capital letters refer to matrices and bold lowercase letters represent vectors. Matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is assumed to have full column rank (i.e., $m > n$); \mathbf{A}_{ij} denotes the (i, j) entry of \mathbf{A} , \mathbf{a}_j is the j^{th} column of \mathbf{A} , and $\mathbf{A}_k \in \mathbb{R}^{n \times k}$ is one of the $\binom{m}{k}$ submatrices of \mathbf{A} . $\mathcal{L}_{\mathbf{A}}$ denotes the subspace spanned by the columns of \mathbf{A} . $\mathbf{P}_{\mathbf{A}}^{\perp} = \mathbf{I} - \mathbf{A}\mathbf{A}^{\dagger}$ is the projection operator onto the orthogonal complement of $\mathcal{L}_{\mathbf{A}}$ where $\mathbf{A}^{\dagger} = (\mathbf{A}^{\top}\mathbf{A})^{-1}\mathbf{A}^{\top}$ is the Moore-Penrose pseudo-inverse of \mathbf{A} and \mathbf{I} is the identity matrix with dimension equal to the number of rows in \mathbf{A} . Likewise, \mathcal{L}_k , \mathbf{A}_k^{\dagger} , and \mathbf{P}_k are similar objects defined in connection to \mathbf{A}_k (here \mathbf{P}_k denotes the projection operator onto \mathcal{L}_k). $\mathcal{I} = \{1, \dots, m\}$ is the set of column indices, $\mathcal{S}_{\text{opt}} = \{1, \dots, k\}$ is the set of indices corresponding to nonzero elements of \mathbf{x} , and \mathcal{S}_i is the set of selected indices at the end of the i^{th} iteration of OLS. For a scalar random variable X , $X \sim \mathcal{B}(\frac{1}{2}, \pm 1)$ denotes that X is a Bernoulli random variable and takes values 1 and -1 with equal probability. For a non-scalar object such as matrix \mathbf{A} , $\mathbf{A} \sim \mathcal{N}(0, \frac{1}{n})$ implies that the entries of \mathbf{A} are drawn independently from a zero-mean Gaussian distribution with variance $\frac{1}{n}$. Similar definitions hold for $\mathbf{A} \sim \mathcal{B}(\frac{1}{2}, \pm \frac{1}{\sqrt{n}})$.

B. The Orthogonal Least-Squares Algorithm

The OLS algorithm sequentially projects columns of \mathbf{A} onto a residual vector and selects the column that leads to the

smallest residual norm. Specifically, in the i^{th} iteration OLS chooses a new index j_s according to

$$j_s = \arg \min_{j \in \mathcal{I} \setminus \mathcal{S}_{i-1}} \left\| \mathbf{y} - \mathbf{A}_{\mathcal{S}_{i-1} \cup \{j\}} \mathbf{A}_{\mathcal{S}_{i-1} \cup \{j\}}^\dagger \mathbf{y} \right\|_2, \quad (3)$$

This procedure is computationally more expensive than OMP since in addition to solving a least-squares problem to update the residual vector, orthogonal projections of the columns of \mathbf{A} need to be found in each step of OLS. Note that the performances of OLS and OMP are identical when the columns of \mathbf{A} are orthogonal.¹ It is worthwhile pointing out the difference between OMP and OLS. In each iteration of OMP, an element most correlated with the current residual is chosen. OLS, on the other hand, selects a column least expressible by previously selected columns which, in turn, minimizes the approximation error.

It was shown in [37] that the index selection criterion 3 can alternatively be expressed as

$$j_s = \arg \max_{j \in \mathcal{I} \setminus \mathcal{S}_{i-1}} \left| \mathbf{r}_{i-1}^\top \frac{\mathbf{P}_{i-1}^\perp \mathbf{a}_j}{\|\mathbf{P}_{i-1}^\perp \mathbf{a}_j\|_2} \right|, \quad (4)$$

where \mathbf{r}_{i-1} denotes the residual vector in the i^{th} iteration. Moreover, projection matrix needed for subsequent iteration is related to the current projection matrix by the following recursion,

$$\mathbf{P}_{i+1}^\perp = \mathbf{P}_i^\perp - \frac{\mathbf{P}_i^\perp \mathbf{a}_{j_s} \mathbf{a}_{j_s}^\top \mathbf{P}_i^\perp}{\|\mathbf{P}_i^\perp \mathbf{a}_{j_s}\|_2^2}. \quad (5)$$

It should be noted that \mathbf{r}_{i-1} in 4 can be replaced by \mathbf{y} because of the idempotent property of the projection matrix, i.e.,

$$\mathbf{P}_i^\perp = \mathbf{P}_i^{\perp \top} = \mathbf{P}_i^{\perp 2}. \quad (6)$$

This substitution reduces complexity of OLS although, when sparsity level k is unknown, the norm of \mathbf{r}_i still needs to be computed since it is typically used when evaluating a stopping criterion. OLS is formalized as Algorithm 1 and referred to in the analysis presented in Section IV-A.

Algorithm 1 Orthogonal Least-Squares (OLS)

Input: \mathbf{y} , \mathbf{A} , sparsity level k

Output: recovered support \mathcal{S}_k , estimated signal $\hat{\mathbf{x}}_k$

Initialize: $\mathcal{S}_0 = \emptyset$, $\mathbf{P}_0^\perp = \mathbf{I}$

for $i = 1$ to k **do**

$$1. j_s = \operatorname{argmax}_{j \in \mathcal{I} \setminus \mathcal{S}_{i-1}} \left| \mathbf{y}^\top \frac{\mathbf{P}_{i-1}^\perp \mathbf{a}_j}{\|\mathbf{P}_{i-1}^\perp \mathbf{a}_j\|_2} \right|$$

$$2. \mathcal{S}_i = \mathcal{S}_{i-1} \cup \{j_s\}$$

$$3. \mathbf{P}_{i+1}^\perp = \mathbf{P}_i^\perp - \frac{\mathbf{P}_i^\perp \mathbf{a}_{j_s} \mathbf{a}_{j_s}^\top \mathbf{P}_i^\perp}{\|\mathbf{P}_i^\perp \mathbf{a}_{j_s}\|_2^2}$$

end for

$$4. \hat{\mathbf{x}}_k = \mathbf{A}_{\mathcal{S}_k}^\dagger \mathbf{y}$$

¹In fact, orthogonality of the columns of \mathbf{A} implies that the objective function in 2 is modular, and hence both methods are optimal when $\nu = 0$.

III. NEW RESULTS ON SPARSE RECOVERY FROM RANDOM MEASUREMENTS VIA OLS

In this section, we first study the performance of OLS in the random measurements and noise-free scenario; specifically, we consider 1 where the elements of \mathbf{A} are drawn from a Gaussian or a Bernoulli distribution and $\nu = 0$, and derive conditions for the exact recovery via OLS. Then we generalize this result to the noisy scenario.

A. Lemmas

We begin by stating several lemmas which will later be used in the proofs of main theorems.

Lemma III.1, a consequence of the convexity of subspaces, establishes that the orthogonal projection onto a subspace is unique regardless of the choice of basis.

Lemma III.1. *Let $\mathcal{F} \subset \mathbb{R}^n$ denote a k -dimensional subspace spanned by the columns of \mathbf{A} and let \mathbf{B} denote a matrix whose columns form an orthonormal basis for \mathcal{F} . Then the orthogonal projection matrices $\mathbf{P}_\mathbf{A}$ and $\mathbf{P}_\mathbf{B}$ are identical, i.e., $\mathbf{P}_\mathbf{A} = \mathbf{P}_\mathbf{B}$.*

Proof. See Appendix A-A. ■

Lemma III.2 states that the projection of a random vector drawn from a Gaussian or a Bernoulli distribution onto a random subspace preserves its Euclidean norm (within a normalizing factor expressed in terms of the problem parameters).

Lemma III.2. *Assume $\mathbf{A} \sim \mathcal{N}(0, 1/n)$ or $\mathbf{A} \sim \mathcal{B}(\frac{1}{2}, \pm \frac{1}{\sqrt{n}})$. Let $\mathbf{A}_k \in \mathbb{R}^{n \times k}$ be a submatrix of \mathbf{A} . Then, $\forall \mathbf{u} \in \mathbb{R}^n$ statistically independent of \mathbf{A}_k drawn according to $\mathbf{u} \sim \mathcal{N}(0, 1/n)$ or $\mathbf{u} \sim \mathcal{B}(\frac{1}{2}, \pm \frac{1}{\sqrt{n}})$, it holds that $\mathbb{E} \|\mathbf{P}_k \mathbf{u}\|_2^2 = \frac{k}{n} \mathbb{E} \|\mathbf{u}\|_2^2$.*

Proof. See Appendix A-B. ■

Lemma III.3, adapted from [38], [39], establishes that projection of a vector drawn from a Gaussian or a Bernoulli distribution onto the column span of Gaussian or Bernoulli matrices is with high probability concentrated around its expected value.²

Lemma III.3. *Assume the conditions stated in Lemma III.2 and let $c_0(\epsilon) = \frac{\epsilon^2}{4} - \frac{\epsilon^3}{6}$. Then,*

$$\Pr\left\{ \left| \|\mathbf{P}_k \mathbf{u}\|_2^2 - \frac{k}{n} \mathbb{E} \|\mathbf{u}\|_2^2 \right| \leq \epsilon \frac{k}{n} \mathbb{E} \|\mathbf{u}\|_2^2 \right\} \geq 1 - 2e^{-kc_0(\epsilon)}. \quad (7)$$

Lemma III.4 states inequalities between maximum and minimum singular values of a matrix and its submatrices.

Lemma III.4. *Let \mathbf{A} , \mathbf{B} , and \mathbf{C} be full rank tall matrices such that $\mathbf{C} = [\mathbf{A}, \mathbf{B}]$. Then*

$$\sigma_{\min}(\mathbf{A}) \geq \sigma_{\min}(\mathbf{C}), \quad \sigma_{\max}(\mathbf{A}) \leq \sigma_{\max}(\mathbf{C}), \quad (8a)$$

$$\sigma_{\min}(\mathbf{B}) \geq \sigma_{\min}(\mathbf{C}), \quad \sigma_{\max}(\mathbf{B}) \leq \sigma_{\max}(\mathbf{C}). \quad (8b)$$

Proof. See Appendix A-C. ■

²It should be noted that the results of Lemma III.2 and III.3 are not limited to Gaussian and Bernoulli random vectors; in fact, similar expressions can be established if \mathbf{u} is a spherically symmetric random vector.

B. Noiseless measurements

The following theorem establishes that when the coefficient matrix consists of entries drawn from a Gaussian or a Bernoulli distribution and the measurements are noise-free, OLS with high probability recovers an unknown sparse vector from the linear combinations of its entries.

Theorem III.5. Suppose $\mathbf{x} \in \mathbb{R}^m$ is an arbitrary sparse vector with $k < m$ non-zero entries. Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ be a random matrix with entries drawn uniformly and independently from either $\mathcal{N}(0, 1/n)$ or $\mathcal{B}(\frac{1}{2}, \pm \frac{1}{\sqrt{n}})$. Let Σ denote the event that given noiseless measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$, OLS can recover \mathbf{x} in k iterations. Then $\Pr\{\Sigma\} \geq p_1 p_2 p_3$, where

$$\begin{aligned} p_1 &= \left(1 - 2e^{-(n-k+1)c_0(\epsilon)}\right)^2, \\ p_2 &= 1 - 2\left(\frac{12}{\delta}\right)^k e^{-nc_0(\frac{\delta}{2})}, \text{ and} \\ p_3 &= \left(1 - 2 \sum_{i=0}^{k-1} e^{-\frac{n}{k-i} \frac{1-\epsilon}{1+\epsilon} (1-\delta)^2}\right)^{m-k}, \end{aligned} \quad (9)$$

for any $0 < \epsilon < 1$ and $0 < \delta < 1$.

Proof. See Appendix B. ■

Using the result of Theorem III.5, one can numerically show that OLS successfully recovers k -sparse \mathbf{x} if the number of measurements is linear in k and logarithmic in m (i.e., the length of \mathbf{x}).

Corollary III.5.1. Let $\mathbf{x} \in \mathbb{R}^m$ be an arbitrary k -sparse vector and let $\mathbf{A} \in \mathbb{R}^{n \times m}$ denote a matrix with entries that are drawn uniformly and independently from either $\mathcal{N}(0, 1/n)$ or $\mathcal{B}(\frac{1}{2}, \pm \frac{1}{\sqrt{n}})$; moreover, assume that $n \geq \max\{\frac{2}{C_1} k \log \frac{m}{\beta}, \frac{C_2 k + \log \frac{12}{\beta^2}}{C_3}\}$, where $0 < \beta < 1$ and C_1, C_2 , and C_3 are positive constants independent of β, n, m , and k . Given noiseless measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$, OLS can recover \mathbf{x} in k iterations with probability of success exceeding $1 - \beta^2$.

Proof. Let us first take a closer look at p_3 . Note that $(1-x)^l \geq 1 - lx$ is valid for $x \leq 1$ and $l \geq 1$; since replacing $k-i$ with k in expression of p_3 in 9 decreases p_3 , $k(m-k) \leq \frac{m^2}{4}$ implies

$$p_3 \geq 1 - \frac{m^2}{2} e^{-C_1 \frac{n}{k}}, \quad (10)$$

where $C_1 = \frac{1-\epsilon}{1+\epsilon}(1-\delta)^2 > 0$. Multiplying both sides of 10 with p_1 and p_2 and discarding some positive higher order terms results in

$$\Pr\{\Sigma\} \geq 1 - \frac{m^2}{2} e^{-C_1 \frac{n}{k}} - 2e^{\log \frac{12}{\delta} k} e^{-nc_0(\frac{\delta}{2})} - 4e^{c_0(\epsilon)k} e^{-nc_0(\epsilon)}. \quad (11)$$

This inequality is readily simplified by defining positive constants $C_2 = \max_{0 < \epsilon, \delta < 1} \{\log \frac{12}{\delta}, c_0(\epsilon)\}$ and $C_3 = \min_{0 < \epsilon, \delta < 1} \{c_0(\frac{\delta}{2}), c_0(\epsilon)\}$,

$$\Pr\{\Sigma\} \geq 1 - \frac{m^2}{2} e^{-C_1 \frac{n}{k}} - 6e^{C_2 k} e^{-nC_3}. \quad (12)$$

We would like to show that $\Pr\{\Sigma\} \geq 1 - \beta^2$. To this end, it suffices to demonstrate that

$$\beta^2 \geq \frac{m^2}{2} e^{-C_1 \frac{n}{k}} + 6e^{C_2 k} e^{-nC_3}. \quad (13)$$

Let $n \geq \frac{C_2 k + \log \frac{12}{\beta^2}}{C_3}$.³ This ensures $6e^{C_2 k} e^{-nC_3} \leq \frac{\beta^2}{2}$ and thus gives the desired result. Moreover,

$$n \geq \max\left\{\frac{2}{C_1} k \log \frac{m}{\beta}, \frac{C_2 k + \log \frac{12}{\beta^2}}{C_3}\right\} \quad (14)$$

is adequate to guarantee $\Pr\{\Sigma\} \geq 1 - \beta^2$ with $0 < \beta < 1$ and $\log \frac{m}{\beta} > 0$ for all $m \geq 1$. ■

Remark 1: Note that when $k \rightarrow \infty$ (and so do m , and n), p_1, p_2 , and p_3 overwhelmingly approach 1. Therefore, one may assume very small ϵ and δ which implies $C_1 \approx 1$.

C. Noisy measurements

We now turn to the general case of noisy random measurements and study the conditions under which OLS exactly recovers the support of \mathbf{x} with high probability.

Theorem III.6. Let $\mathbf{x} \in \mathbb{R}^m$ be an arbitrary k -sparse vector and let $\mathbf{A} \in \mathbb{R}^{n \times m}$ denote a matrix with entries that are drawn uniformly and independently from either $\mathcal{N}(0, 1/n)$ or $\mathcal{B}(\frac{1}{2}, \pm \frac{1}{\sqrt{n}})$. Given the noisy measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\nu}$ where $\|\boldsymbol{\nu}\|_2 \leq \epsilon_\nu$, and $\boldsymbol{\nu}$ is independent of \mathbf{A} and \mathbf{x} , if $\min_{\mathbf{x}_j \neq 0} |\mathbf{x}_j| \geq (1 + \delta + t)\epsilon_\nu$ for any $t > 0$, OLS can recover \mathbf{x} in k iterations with probability of success $\Pr\{\Sigma\} \geq p_1 p_2 p_3$ where,

$$\begin{aligned} p_1 &= \left(1 - 2e^{-(n-k+1)c_0(\epsilon)}\right)^2 \\ p_2 &= 1 - 2\left(\frac{12}{\delta}\right)^k e^{-nc_0(\frac{\delta}{2})}, \text{ and} \\ p_3 &= \left(1 - 2 \sum_{i=0}^{k-1} e^{-\frac{n}{k-i} \frac{1-\epsilon}{1+\epsilon} (1-\delta)^4} \frac{1}{k \left[\frac{1}{(k-i)t^2} + (1+\delta)^2\right]}\right)^{m-k} \end{aligned} \quad (15)$$

for any $0 < \epsilon < 1, 0 < \delta < 1$.

Proof. See Appendix C. ■

Remark 2: If we define $\text{SNR} = \frac{\|\mathbf{A}\mathbf{x}\|_2^2}{\|\boldsymbol{\nu}\|_2^2}$, the condition $\min_{\mathbf{x}_j \neq 0} |\mathbf{x}_j| \geq (1 + \delta + t)\epsilon_\nu$ implies

$$\text{SNR} \approx k(1 + \delta + t)^2, \quad (16)$$

which suggests that for exact support recovery via OLS, SNR should scale linearly with sparsity level.

Corollary III.6.1. Let $\mathbf{x} \in \mathbb{R}^m$ be an arbitrary k -sparse vector and let $\mathbf{A} \in \mathbb{R}^{n \times m}$ denote a matrix with entries that are drawn uniformly and independently from either $\mathcal{N}(0, 1/n)$ or $\mathcal{B}(\frac{1}{2}, \pm \frac{1}{\sqrt{n}})$; moreover, assume that $n \geq \max\{\frac{2}{C_1} k \log \frac{m}{\beta}, \frac{C_2 k + \log \frac{12}{\beta^2}}{C_3}\}$ where $0 < \beta < 1$ and C_1, C_2 , and C_3 are positive constants that are independent of β, n, m , and k . Given the noisy measurements $\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\nu}$ where $\boldsymbol{\nu} \sim \mathcal{N}(0, \sigma^2)$ is independent of \mathbf{A} and \mathbf{x} , if $\min_{\mathbf{x}_j \neq 0} |\mathbf{x}_j| \geq C_4 \|\boldsymbol{\nu}\|_2$ for some $C_4 > 1$, OLS can recover \mathbf{x} in k iterations with probability of success exceeding $1 - \beta^2$.

Proof. The proof follows the steps similar to those of Corollary III.5.1, leading to constants $C_1 = \frac{1-\epsilon}{1+\epsilon}(1-\delta)^4(1 +$

³This guarantees $n \geq k$ for all m, n , and k .

$t^2(1 + \delta)^2)^{-1}$, $C_2 = \max_{0 < \epsilon, \delta < 1} \{\log \frac{12}{\delta}, c_0(\epsilon)\} > 0$, $C_3 = \min_{0 < \epsilon, \delta < 1} \{c_0(\frac{\delta}{2}), c_0(\epsilon)\} > 0$, and $C_4 = (1 + \delta + t)$. ■

Remark 3: In general, for the case of noisy measurements C_1 is smaller than that of the noiseless setup, resulting in a more restrictive sampling requirement for the former.

IV. ACCELERATED SCHEMES FOR SPARSE RECOVERY

The complexity of OLS is dominated by the so-called identification and update steps, formalized as steps 1 and 3 of Algorithm 1 in Section II, respectively; in these steps, the algorithm evaluates projections $\mathbf{P}_{i-1}^\perp \mathbf{a}_j$ of remaining columns onto the space spanned by the selected ones and then compute the projection matrix \mathbf{P}_i needed for the next iteration. This may be practically infeasible in applications that involve dealing with high-dimensional data. To this end, we here establish a set of recursions which further reduce the complexity of the identification and update steps.

Theorem IV.1. *Let \mathbf{r}_i denote the residual vector in the i^{th} iteration of OLS (the algorithm is initialized with $\mathbf{r}_0 = \mathbf{y}$). The identification step (labeled as step 1 in Algorithm 1) in the $(i + 1)^{\text{st}}$ iteration of OLS can be rephrased as*

$$j_s = \operatorname{argmax}_{j \in \mathcal{I}} \|\mathbf{z}_j \mathbf{q}_j\|_2, \quad (17)$$

where

$$\mathbf{z}_j = \mathbf{r}_i^\top \mathbf{a}_j, \quad \mathbf{t} = \mathbf{a}_j - \sum_{l=1}^i \frac{\mathbf{a}_j^\top \mathbf{u}_l}{\|\mathbf{u}_l\|_2^2} \mathbf{u}_l, \quad \mathbf{q}_j = \frac{1}{\mathbf{a}_j^\top \mathbf{t}} \mathbf{t}. \quad (18)$$

Furthermore, the residual vector \mathbf{r}_{i+1} required for the next iteration is formed as

$$\mathbf{u}_{i+1} = \mathbf{q}_{j_s}, \quad \mathbf{r}_{i+1} = \mathbf{r}_i - z_{j_s} \mathbf{u}_{i+1}. \quad (19)$$

Proof. See Appendix D. ■

Note that the set of recursive equations established in Theorem IV.1 complies with the geometric interpretation of OLS. In particular, after orthogonalizing the subset of selected columns, OLS identifies a new column and adds it to the subspace, thus expanding it. This point is formalized as Corollary IV.1.1.

Corollary IV.1.1. *Let $\{\tilde{\mathbf{a}}_l\}_{l=1}^i$ denote the set of columns selected in the first i iterations of the OLS algorithm and let $\mathcal{L} = \{\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_i\}$ be the subspace spanned by these columns. Then $\{\mathbf{u}_l\}_{l=1}^i$ generated according to Theorem IV.1 forms an orthogonal basis for \mathcal{L}_i .*

A. Accelerated OLS

Using the relations established in Theorem IV.1, we propose a computationally efficient modification of the OLS algorithm referred to as Accelerated OLS (AOLS) and formalize it as Algorithm 2. It is straightforward to see that the computational costs of Algorithm 1 and Algorithm 2 are $\mathcal{O}(mn^2k)$ and $\mathcal{O}(mnk^2)$, respectively. Clearly, accelerated OLS is less complex than the conventional OLS in a variety of practical scenarios since, typically, $k \ll n$. However, due to constants and lower order terms in the complexity expressions,

Algorithm 2 Accelerated Orthogonal Least-Squares (AOLS)

Input: \mathbf{y} , \mathbf{A} , sparsity level k
Output: recovered support \mathcal{S}_k , estimated signal $\hat{\mathbf{x}}_k$
Initialize: $\mathcal{S}_0 = \emptyset$, $\mathbf{r}_0 = \mathbf{y}$
for $i = 1$ to k **do**
 for $j \in \mathcal{I} \setminus \mathcal{S}_{i-1}$ **do**
 1. $\mathbf{z}_j = \mathbf{r}_{i-1}^\top \mathbf{a}_j$
 2. $\mathbf{t} = \mathbf{a}_j - \sum_{l=1}^{i-1} \frac{\mathbf{a}_j^\top \mathbf{u}_l}{\|\mathbf{u}_l\|_2^2} \mathbf{u}_l$
 3. $\mathbf{q}_j = \frac{1}{\mathbf{a}_j^\top \mathbf{t}} \mathbf{t}$
 end for
 4. $j_s = \operatorname{argmax}_{j \in \mathcal{I} \setminus \mathcal{S}_{i-1}} \|\mathbf{z}_j \mathbf{q}_j\|_2$
 5. $\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{j_s\}$
 6. $\mathbf{u}_i = \mathbf{q}_{j_s}$, $\mathbf{r}_i = \mathbf{r}_{i-1} - z_{j_s} \mathbf{u}_i$
end for
7. $\hat{\mathbf{x}}_k = \mathbf{A}_{\mathcal{S}_k}^\dagger \mathbf{y}$

in different applications and for varied dimensions of the problem, one implementation may be preferred over the other. To this end, we carefully analyze computational complexity of AOLS and compare it precisely to that of OLS. Note that in the following discussion we assume scalar-vector, vector-vector, and matrix-vector multiplications require n , $2n$, and $2n^2$ operations, respectively (here, the dimensions of said vector and matrix are n and $n \times n$, respectively).

Consider the first iteration of Algorithm 1. The first step requires $(2n^2 + 4n)m$ operations as one needs to compute a matrix-vector and two vector-vector multiplications for each of m columns. Step 3 requires $\frac{7}{2}n^2 + \frac{5}{2}n$ operations. To see that, note that the computation of $\mathbf{P}_i^\perp \mathbf{a}_{j_s}$ requires $2n^2$ operations, symmetric matrix $\mathbf{P}_i^\perp \mathbf{a}_{j_s} \mathbf{a}_{j_s}^\top \mathbf{P}_i^\perp$ requires $\frac{n(n+1)}{2}$ operations, $\|\mathbf{P}_i^\perp \mathbf{a}_{j_s}\|_2$ can be computed by a trace operator with n operations and requires multiplication involving $\frac{n(n+1)}{2}$ entries. Finally, it takes $\frac{n(n+1)}{2}$ operations to evaluate the projection matrix used in the next iteration. Therefore, the first iteration requires in total $(2n^2 + 5n)m + (\frac{7}{2}n^2 + \frac{5}{2}n)$ operations. Since step 1 is performed only on the columns that are not yet selected, computational cost of the i^{th} iteration is $(2n^2 + 4n)(m - i + 1) + (\frac{7}{2}n^2 + \frac{5}{2}n)$. In conclusion, the total number of operations required by OLS is $(2n^2 + 4n) \left(km - \frac{k(k-1)}{2} \right) + k(\frac{7}{2}n^2 + \frac{5}{2}n)$.

The complexity of Accelerated OLS is examined next. By the i^{th} iteration, $i - 1$ columns have already been chosen. In the inner loop, step 1 and step 3 require $2n$ and $3n$ operations, respectively. Step 2 is computed only for $i > 1$ and costs $6n(i - 1) - 2n(i - 2) = 2n(2i - 1)$ since the norm of vectors \mathbf{u}_l can be stored inexpensively and used in subsequent iterations. Step 4 requires $2n(m - i + 1)$, and step 6 needs $2n$ operations. Thus, the aggregate complexity of Algorithm 2 is $7n \left(km - \frac{k(k-1)}{2} \right) + 2nk + 2n \sum_{i=1}^k (m - i + 1)(2i - 1)$. The result of this comparison are summarized in Table I. To

Table I. Computational Complexity of OLS and Accelerated OLS

Algorithm	Number of arithmetic operations
OLS	$4n \left(km - \frac{k(k-1)}{2} \right) + \frac{5}{2}nk + 2n^2 \left(km - \frac{k(k-1)}{2} \right) + \frac{7}{2}n^2k$
Accelerated OLS	$5n \left(km - \frac{k(k-1)}{2} \right) + 2nk + 2nk(k+1)(m+1) - \frac{2}{3}k(k+1)(2k+1)$

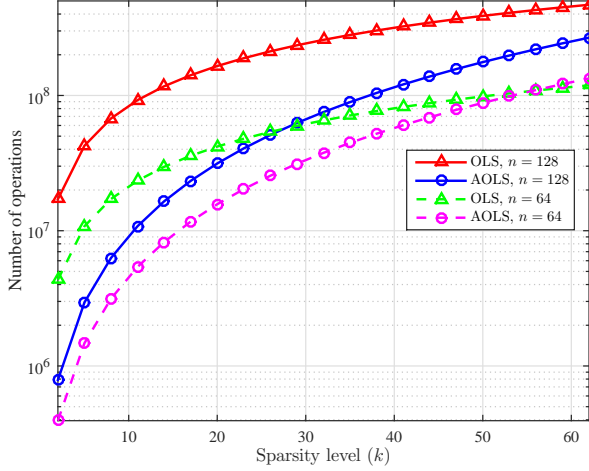
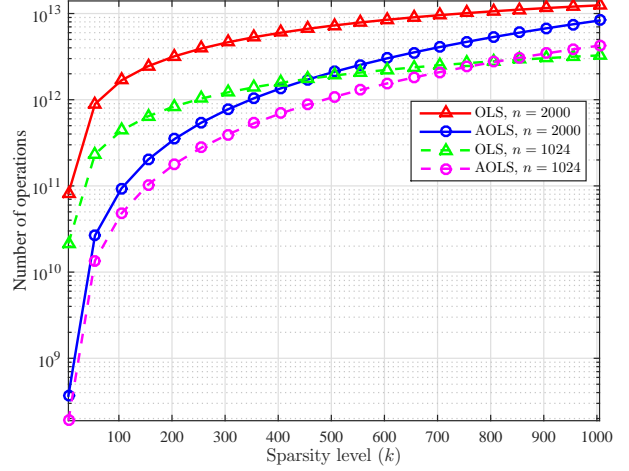
(a) $m = 256$ (b) $m = 2048$

Fig. 1. Required number of operations for OLS and Accelerated OLS.

illustrate how the complexities of OLS and AOLS compare, we consider 4 different scenarios. In the first two, dimension of the unknown vector is $m = 256$ while we assume a relatively low number of measurements, $n = 64$ and $n = 128$. The sparsity level is varied to cover a variety of scenarios. We plot the number of operations needed for OLS and Accelerated OLS in this setting in Fig. 1(a). It can be observed that when $n = 128$, the complexity of performing AOLS is uniformly smaller than that of OLS (approximately 10 times smaller for the relatively sparse vectors). When the number of acquired measurements decreases to $n = 64$, for $k = 53$ and higher AOLS costs more; however, this is not an interesting setting since many of sparse recovery algorithms, including OLS and AOLS, perform poorly therein. We also consider the high-dimensional case of $m = 2048$ with relatively large number of measurements $n = 1024$ and $n = 2000$ and plot the resulting complexity expressions in Fig. 1 (b). For all practical scenarios, AOLS requires fewer operations to perform support recovery. In fact, for relatively high n and low k , the cost of doing AOLS is approximately 100 times smaller than that of OLS (while their accuracies, of course, coincide).

B. Generalized OLS

The AOLS algorithm from Section IV-A is a computationally more efficient implementation of orthogonal least-squares, gaining speed over traditional OLS without sacrificing accuracy. In this subsection, we propose an extension of OLS, Generalized OLS (GOLS), which provides higher accuracy at often lower speeds compared to OLS. In particular, GOLS selects multiple (say, L) columns of \mathbf{A} in each step rather than

choosing a single column, ultimately replacing the underdetermined $n \times m$ system of equations by an overdetermined $Lk \times k$ one. This strategy is motivated by the observation that the candidate columns whose projection onto the space orthogonal to that spanned by the previously selected columns is strongly correlated with the observation vector but not chosen in the current step of OLS will likely be selected in subsequent steps of the algorithm; therefore, selecting several “good” candidates in each step accelerates the selection procedure and enables sparse reconstruction in fewer steps (and, therefore, requires fewer calculations of the mutual correlations needed to perform the selection). More specifically, the proposed GOLS algorithm performs the following: in each step, the algorithm selects L columns of matrix \mathbf{A} such that their normalized projections onto the orthogonal complement of the subspace spanned by the previously chosen columns have the highest correlation with the residual vector among the non-selected columns. After such columns are identified, we update the orthogonal projection matrix by repeatedly applying 5 L times. For an accelerated implementation of this strategy, we employ 19 to repeatedly generate residual vectors required for consecutive iterations. We continue until a stopping criterion, e.g., a predetermined threshold on the norm of the residual vector is met. Note that the complexity of performing GOLS is typically much lower than that of the conventional OLS; this is due to high probability of finding the true columns using fewer iterations than k and thus reducing the number of time computationally dominant step 1 of Algorithm 1 is performed by GOLS. GOLS is formalized as Algorithm 3.

Algorithm 3 Generalized Orthogonal Least-Squares

Input: \mathbf{y} , \mathbf{A} , sparsity level k , threshold ϵ
Output: recovered support \mathcal{S}_k , estimated signal $\hat{\mathbf{x}}_k$
Initialize: $\mathcal{S}_0 = \emptyset$, $\mathbf{r}_0 = 0$
while $\|\mathbf{r}_i\|_2 \geq \epsilon$ **do**
 1. Select $\{i_{s_1}, \dots, i_{s_L}\}$ corresponding to L largest terms:
 $\left| \mathbf{y}^\top \frac{\mathbf{P}_{i-1}^\perp \mathbf{a}_j}{\|\mathbf{P}_{i-1}^\perp \mathbf{a}_j\|_2} \right|$ for OLS or $\|z_j \mathbf{q}_j\|_2$ for AOLS
 2. $\mathcal{S}_i = \mathcal{S}_{i-1} \cup \{i_{s_1}, \dots, i_{s_L}\}$
 3. Update \mathbf{P}_i^\perp using 5 for OLS or update \mathbf{r}_i using 19 for AOLS
end while
 4. $\hat{\mathbf{x}}_k = \mathbf{A}_{\mathcal{S}_k}^\dagger \mathbf{y}$

V. SIMULATIONS

A. Recovery with random measurements

Here we verify our theoretical results by comparing them to the empirical ones obtained via Monte Carlo simulations. In each trial, we select locations of the nonzero elements of \mathbf{x} uniformly at random and draw them from a normal distribution. Entries of the coefficient matrix \mathbf{A} are also generated at random according to $\mathcal{N}(0, \frac{1}{n})$. Fig. 2 plots the number of noiseless measurement n needed to achieve at least 0.95 probability of perfect recovery as a function of the sparsity level k . The length of the unknown vector \mathbf{x} here is set to $m = 256$, and the results (shown as circles) are averaged over 1000 independent trials. The solid regression line in Fig. 2 implies linear relation between n and k as predicted by Corollary III.5.1. Specifically, for the considered setting, $n \approx 0.75 k \log m$. Note that according to Remark 1, for a high dimensional problem, $C_1 \approx 1$ and one would expect $n \geq 2 k \log m$ for all m , and k which clearly is not the case in the setting examined here. Consequently, Fig. 2 demonstrates that our theoretical result is slightly pessimistic which is as a results of estimates that we employed in the proof of Corollary III.5.1. Fig. 3 illustrates performance of OLS in the presence of measurement noise, assumed to be additive white Gaussian (AWGN). Here we fix the number of measurement $m = 256$ and vary k . The SNR is adjusted linearly according to sparsity level. Specifically, in each trail we determine variance of ν such that $\text{SNR} = \|\mathbf{A}\mathbf{x}\|_2 / \|\nu\|_2 = 100 k$. We plot the empirically determined minimum number of measurements (shown as circles) required for at least 0.9 exact recovery rate. The solid regression line in Fig. 3 implies linear relation between n and k – same as in the noiseless setup except that the slope of the regression line is now steeper. This observation complies with our theoretical results as C_1 in Corollary III.5.1 is larger than C_1 in Corollary III.6.1.

B. Accelerated OLS

To compare the computational complexity of OLS and AOLS, we conduct two experiments. In the first, we set $m = 1024$, $n = 128$, and vary k , corresponding to a problem

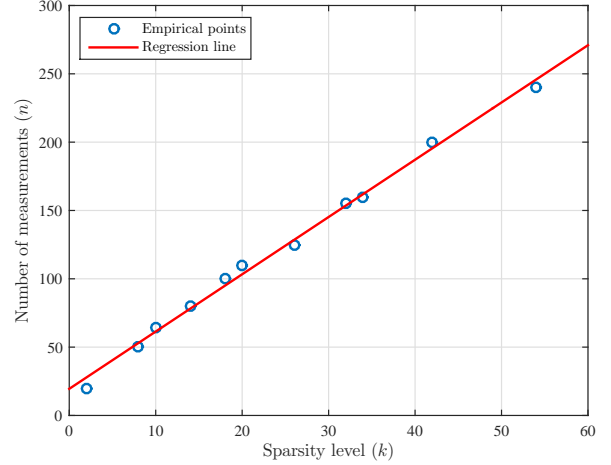


Fig. 2. Number of noiseless measurements required for sparse reconstruction with probability of success at least 95% when $m = 256$. The regression line is $n = 0.7558 k \log m + 19.4798$ with the coefficient of determination $R^2 = 0.9949$.

with highly sparse unknown vector and a relatively low number of measurements. The positions of nonzero elements of \mathbf{x} are chosen uniformly at random and their values drawn from $\mathcal{N}(0, 1)$; the coefficient matrix \mathbf{A} consists of entries drawn from $\mathcal{N}(0, \frac{1}{n})$. The number of elementary operations (flops) of OLS and AOLS is averaged over 1000 Monte Carlo runs emulating noiseless measurement scenario. The results are illustrated in Fig. 4 and compared with the theoretical results. As seen there, AOLS requires significantly smaller computation cost than the conventional OLS. However, as k increases the complexity gap shrinks as predicted by the results of our analysis presented in Section IV-A. In the second experiment, n , m , and k are varied while keeping $n = \frac{m}{2}$ and $k = \sqrt{m}$. We generate \mathbf{x} , \mathbf{A} , and \mathbf{y} in the same way as in the previous experiment. The average flop counts for OLS and AOLS shown in Fig. 8 imply that the computational savings of AOLS over OLS increase significantly as the dimension of the sparse reconstruction problem grows.

C. Generalized OLS

To evaluate performance of the generalized OLS (GOLS) algorithm, we compare it with that of four other sparse recovery algorithms as a function of the sparsity level k . In particular, we considered OMP, OLS, l_1 -norm minimization [12], and Least Absolute Shrinkage and Selection Operator (LASSO) [40]. As typically done in benchmarking tests [17], [19], we used CVX [41] to implement the l_1 minimization and LASSO. We explored various values of L (specifically, $L = 2, 4, 6$) to better understand its effect on the performance of GOLS. The coefficient matrix \mathbf{A} comprises entries drawn at random from $\mathcal{N}(0, \frac{1}{n})^4$. Three different scenarios are considered: (1) the non-zero elements of \mathbf{x} are independent and identically distributed normal random variables, (2) \mathbf{x} is a sparse 2-level

⁴We here omit the setting where $\mathbf{A} \sim \mathcal{B}(\frac{1}{2}, \pm 1)$ as those results are nearly identical to the ones for \mathbf{A} with Gaussian entries.

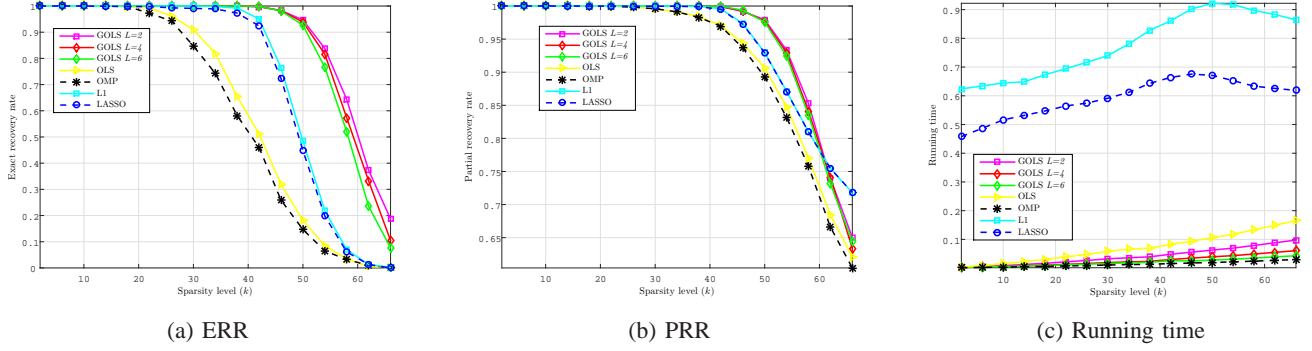


Fig. 6. Performance comparison of GOLDS, OLS, OMP, l_1 -norm minimization and LASSO for $n = 128$, $m = 256$, \mathbf{A} having Gaussian $\mathcal{N}(0, 1/n)$ entries, and the k non-zero components of \mathbf{x} drawn from $\mathcal{N}(0, 1)$ distribution.

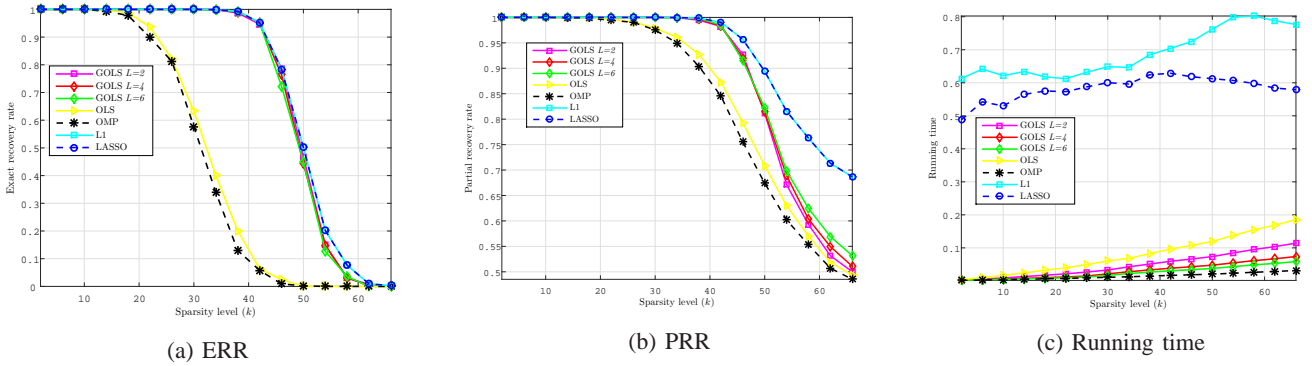


Fig. 7. Performance comparison of GOLDS, OLS, OMP, l_1 -norm minimization and LASSO for $n = 128$, $m = 256$, \mathbf{A} having Gaussian $\mathcal{N}(0, 1/n)$ entries, and the k non-zero components of \mathbf{x} drawn uniformly from $\{\pm 1, \pm 3\}$.

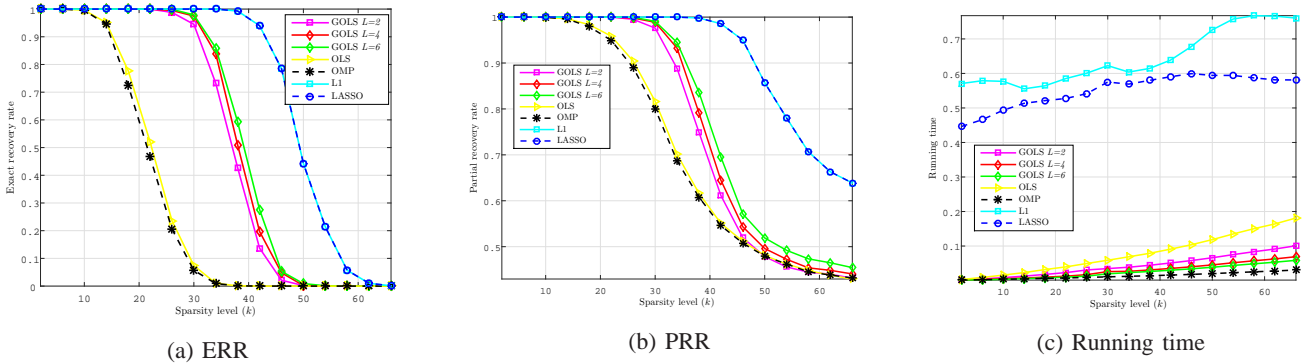


Fig. 8. Performance comparison of GOLDS, OLS, OMP, l_1 -norm minimization and LASSO for $n = 128$, $m = 256$, \mathbf{A} having Gaussian $\mathcal{N}(0, 1/n)$ entries, and the k non-zero components of \mathbf{x} randomly and equally likely set to 1 or -1 .

double polarity pulse-amplitude modulation (PAM) signal, i.e., the nonzero elements of \mathbf{x} are drawn uniformly from the alphabet $\{\pm 1, \pm 3\}$, and (3) \mathbf{x} is a sparse 2-level PAM signal with non-zero components drawn from $\mathcal{B}(\frac{1}{2}, \pm 1)$. In all the considered settings, the locations of non-zero entries of \mathbf{x} are drawn uniformly at random. The number of equations is $n = 128$, the dimension of \mathbf{x} is $m = 256$; the experiment is repeated 1000 times. Performance of each algorithm is characterized by three metrics: (i) Exact Recovery Rate (ERR), defined as the percentage of instances where the support of \mathbf{x} is recovered exactly, (ii) Partial Recovery Rate (PRR),

measuring the fraction of support which is recovered correctly, and (iii) the running time of the algorithm. The results for normally distributed \mathbf{x} are illustrated in Fig. 6. Fig. 7 shows the performance of the algorithms for the 4-level PAM unknown signal while Fig. 8 corresponds to \mathbf{x} being a sparse 2-level PAM signal. As can be seen from Fig. 6, GOLDS outperforms all the competing methods in terms of the exact recovery rate for all values of L . For $k < 62$, GOLDS has the best performance in terms of the partial recovery rate. Moreover, the runtimes of GOLDS are 2nd only to OMP but the accuracy of the latter is significantly worse than that of GOLDS. In

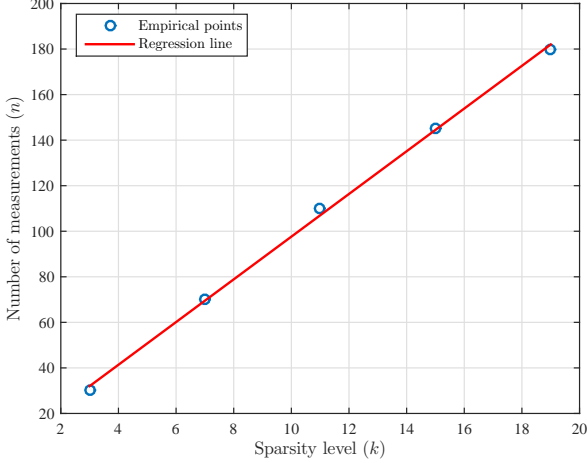


Fig. 3. Number of noisy measurements required for sparse reconstruction with probability of success at least 90% when $m = 256$ and $\text{SNR} = 100k$. The regression line is $n = 1.6907 k \log m + 3.8750$ with the coefficient of determination $R^2 = 0.9988$.

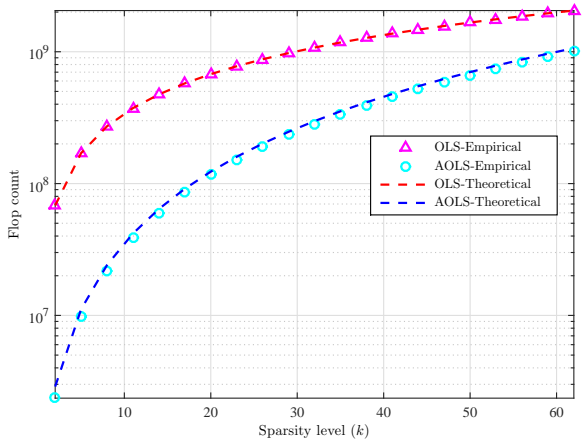


Fig. 4. Average flop (operation) count of OLS and AOLS versus k for $n = 128$, and $m = 1024$.

addition, due to fewest required iterations, GOLS with $L = 6$ is faster than GOLS with $L = 2$ and $L = 4$. For the 4-level PAM signal, Fig. 7 shows that the performance of GOLS is nearly identical to those of the l_1 -norm and LASSO in terms of the exact recovery rate while the l_1 -norm and LASSO deliver higher partial recovery rate at a significantly slower speed. In the case of \mathbf{x} with non-zero entries from $\{+1, -1\}$ studied in Fig. 8, l_1 -norm/LASSO methods perform the best (and are the slowest) while the GOLS offers reasonably accurate performance at a relatively high speed. The results presented here suggest that the performance of OMP/OLS and hence GOLS is better when there is a decaying relation between the nonzero elements of \mathbf{x} which is consistent with the theoretical results of [35].

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we showed that for normalized Gaussian and Bernoulli coefficient matrices, Orthogonal Least-Squares

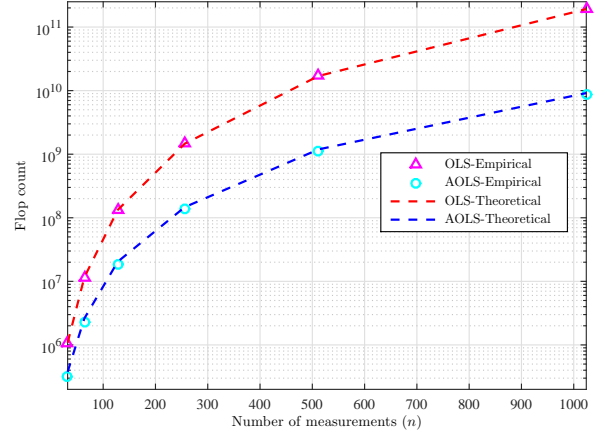


Fig. 5. Average flop (operation) count of OLS and AOLS versus k where m varies from 64 to 2048, $n = m/2$, and $k = \lfloor \sqrt{m} \rfloor$.

(OLS) with high probability recovers m -dimensional sparse signals with no more than k non-zero entries in at most k iterations from $\mathcal{O}(k \log m)$ noiseless random linear measurements. We extended this result to the scenario where the measurements are perturbed with l_2 -bounded noise. Specifically, we if the non-zero elements of an unknown vector are sufficiently large, $\mathcal{O}(k \log m)$ random linear measurements is sufficient to guarantee recovery with high probability. Simulation results demonstrate that $\mathcal{O}(k \log m)$ measurement is indeed sufficient for sparse reconstruction that is exact with probability arbitrarily close to one. In addition, we derived a set of expressions which facilitate efficient recursive updates of the orthogonal projection operator and the computation of the residual vector by employing only linear equations; this has led to a fast variant of OLS suitable for high-dimensional applications which we refer to as accelerated OLS. Moreover, we introduced a novel generalized OLS for sparse linear regression that forms the subset of columns of a coefficient matrix in an underdetermined system of equations by sequentially selecting multiple candidate columns. Since multiple indices are selected without any additional cost, the running time of the algorithm is effectively reduced as compared to OLS. Generalized OLS is more favorable than convex optimization based methods whose complexity grows faster with the dimension of the problem, i.e., n and m . Simulation studies demonstrate that generalized OLS outperforms competing greedy methods, OLS and OMP, while being computationally more efficient than l_1 -norm minimization and LASSO.

As part of future work, it would be beneficial to establish sampling requirements for random frequency measurements that can be stored and processed efficiently. However, since in this application the columns of the measurement matrix are no longer statistically independent, it appears difficult to analyze iterations of OLS. A way forward may be to employ Restricted Isometry Property (RIP) of the random frequency matrix. Moreover, in some cases only partial support recovery is needed and it is of interest to establish analysis framework for such settings that is analogous to the one in the current paper.

In addition, It would be valuable to further extend the analysis carried out in Section III to find sampling requirements for exact recovery of generalized OLS.

APPENDIX A PROOF OF LEMMAS

A. Proof of Lemma III.1

Let $\{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ be an orthonormal basis for \mathcal{F} . Select $\{\mathbf{b}_{k+1}, \dots, \mathbf{b}_n\}$ such that $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ form an orthonormal basis for \mathbb{R}^n . Since every vector in \mathcal{F} projects onto itself, $\mathbf{P}_A \mathbf{b}_i = \mathbf{P}_B \mathbf{b}_i = \mathbf{b}_i, \forall i \in \{1, \dots, k\}$. Moreover, since \mathcal{F} is k dimensional and $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ is a collection of orthonormal basis vectors, $\mathbf{P}_A \mathbf{b}_j = \mathbf{P}_B \mathbf{b}_j = \mathbf{0}, \forall j \in \{k+1, \dots, n\}$. For all $\mathbf{u} \in \mathbb{R}^n$, there exists $\{\beta_i\}_{i=1}^n \in \mathbb{R}$ such that $\mathbf{u} = \sum_{i=1}^n \beta_i \mathbf{b}_i$. It then follows that $\mathbf{P}_A \mathbf{u} = \mathbf{P}_B \mathbf{u} = \sum_{i=1}^k \beta_i \mathbf{b}_i$, and therefore $\mathbf{P}_A = \mathbf{P}_B$, which is the desired result.

B. Proof of Lemma III.2

Recall that \mathbf{P}_k is an orthogonal projection operator for a k -dimensional subspace \mathcal{L}_k spanned by the columns of \mathbf{A}_k . Let $\mathcal{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_k\}$ denote an orthonormal basis for \mathcal{L}_k . There exist a rotation operator \mathcal{R} such that $\mathcal{R}(\mathcal{B}) = \{\mathbf{e}_1, \dots, \mathbf{e}_k\}$, where \mathbf{e}_i is the i^{th} standard unit vector. Hence, the projection of \mathbf{u} onto \mathcal{L}_k is determined by identifying the first k components of $\mathcal{R}(\mathbf{u})$. First, let us consider the case $\mathbf{u} \sim \mathcal{N}(0, 1/n)$. Since a multivariate Gaussian distribution is spherically symmetric, distribution of \mathbf{u} remains unchanged under rotation, i.e., $\mathcal{R}(\mathbf{u}) \sim \mathcal{N}(0, 1/n)$. In the case of a Bernoulli \mathbf{u} , $\mathcal{R}(\mathbf{u})$ is a unit length vector since \mathbf{u} is a unit length vector. Therefore, for both cases it holds that $\mathbb{E} \|\mathcal{R}(\mathbf{u})\|_2 = \mathbb{E} \|\mathbf{u}\|_2$. It follows from the i.i.d. assumption and the linearity of expectation that $\mathbb{E} \|\mathbf{P}_k \mathbf{u}\|_2^2 = \frac{k}{n} \mathbb{E} \|\mathbf{u}\|_2^2$, which completes the proof.

C. Proof of Lemma III.4

We only prove 8a since the roles of \mathbf{A} and \mathbf{B} are interchangeable. Consider $\mathbf{C}^\top \mathbf{C}$. By the definition of a singular value and the fact that $\mathbf{C}^\top \mathbf{C}$ is a positive semidefinite matrix,

$$\mathbf{C}^\top \mathbf{C} = \begin{bmatrix} \mathbf{A}^\top \mathbf{A} & \mathbf{A}^\top \mathbf{B} \\ \mathbf{B}^\top \mathbf{A} & \mathbf{B}^\top \mathbf{B} \end{bmatrix} \preceq \sigma_{\max}(\mathbf{C}^\top \mathbf{C}) \mathbf{I} = \sigma_{\max}^2(\mathbf{C}) \mathbf{I}. \quad (20)$$

Therefore,

$$\mathbf{H}_C = \sigma_{\max}^2(\mathbf{C}) \mathbf{I} - \begin{bmatrix} \mathbf{A}^\top \mathbf{A} & \mathbf{A}^\top \mathbf{B} \\ \mathbf{B}^\top \mathbf{A} & \mathbf{B}^\top \mathbf{B} \end{bmatrix} \succeq 0. \quad (21)$$

Since \mathbf{H}_C is positive semidefinite, Sylvester's criterion implies that the principal minors of \mathbf{H}_C are nonnegative. Therefore, it is clear that $\mathbf{H}_A = \sigma_{\max}^2(\mathbf{C}) \mathbf{I} - \mathbf{A}^\top \mathbf{A} \succeq 0$ since the set of principal minors of \mathbf{H}_A is a subset of the set of principal minors of \mathbf{H}_C . Consider the singular value decomposition $\mathbf{A}^\top \mathbf{A} = \mathbf{U} \mathbf{D} \mathbf{U}^\top$, where \mathbf{U} is a unitary matrix and \mathbf{D} is a diagonal matrix having singular values of $\mathbf{A}^\top \mathbf{A}$ on the main diagonal. Now,

$$\begin{aligned} \mathbf{U}^\top \mathbf{H}_A \mathbf{U} &= \mathbf{U}^\top (\sigma_{\max}^2(\mathbf{C}) \mathbf{I} - \mathbf{A}^\top \mathbf{A}) \mathbf{U} \\ &= \sigma_{\max}^2(\mathbf{C}) \mathbf{U}^\top \mathbf{U} - \mathbf{U}^\top \mathbf{U} \mathbf{D} \mathbf{U}^\top \mathbf{U} \\ &\stackrel{(a)}{=} \sigma_{\max}^2(\mathbf{C}) \mathbf{I} - \mathbf{D} \stackrel{(b)}{\succeq} 0, \end{aligned} \quad (22)$$

where (a) holds because \mathbf{U} is unitary and (b) stems from the fact that \mathbf{H}_A is positive semidefinite and $\mathbf{U}^\top \mathbf{H}_A \mathbf{U}$ is in quadratic form. Since the singular values of \mathbf{A} are square roots of those of $\mathbf{A}^\top \mathbf{A}$, $\sigma_{\max}(\mathbf{A}) \leq \sigma_{\max}(\mathbf{C})$. Following similar reasoning, $\sigma_{\min}(\mathbf{A}) \geq \sigma_{\min}(\mathbf{C})$, which completes the proof.

APPENDIX B PROOF OF THEOREM III.5

Assume, without a loss of generality, that the nonzero components of \mathbf{x} are in the first k locations. This implies that \mathbf{A} can be written as $\mathbf{A} = [\tilde{\mathbf{A}}, \tilde{\tilde{\mathbf{A}}}]$, where $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times k}$ has columns with indices in \mathcal{S}_{opt} and $\tilde{\tilde{\mathbf{A}}} \in \mathbb{R}^{n \times (m-k)}$ has columns with indices in $\mathcal{I} \setminus \mathcal{S}_{\text{opt}}$. For $\mathcal{T}_1 \subset \mathcal{I}$ and $\mathcal{T}_2 \subset \mathcal{I}$, define

$$\mathbf{b}_{j_{\mathcal{T}_1}}^\perp = \frac{\mathbf{a}_j}{\|\mathbf{P}_{\mathcal{T}_1}^\perp \mathbf{a}_j\|_2}, \quad j \in \mathcal{T}_2, \quad (23)$$

where $\mathbf{P}_{\mathcal{T}_1}^\perp$ denotes the projection matrix onto the orthogonal complement of the subspace spanned by the columns of \mathbf{A} with indices in \mathcal{T}_1 . Using the notation of 23, 4 becomes

$$j_s = \arg \max_{j \in \mathcal{I} \setminus \mathcal{S}_{i-1}} \left| \mathbf{r}_{i-1}^\top \mathbf{b}_j^{\mathcal{S}_{i-1}} \right|. \quad (24)$$

In addition, let $\Phi_{\mathcal{S}_i} = [\mathbf{b}_j^{\mathcal{S}_i}] \in \mathbb{R}^{n \times (k-i)}$, $j \in \mathcal{S}_{\text{opt}} \setminus \mathcal{S}_i$, and $\Psi_{\mathcal{S}_i} = [\mathbf{b}_j^{\mathcal{S}_i}] \in \mathbb{R}^{n \times (m-k)}$, $j \in \mathcal{I} \setminus \mathcal{S}_{\text{opt}}$. Assume that in the first i iterations OLS has selected columns from \mathcal{S}_{opt} . According to the selection rule in 24,

$$\rho(\mathbf{r}_i) = \frac{\|\Psi_{\mathcal{S}_i}^\top \mathbf{r}_i\|_\infty}{\|\Phi_{\mathcal{S}_i}^\top \mathbf{r}_i\|_\infty} < 1 \quad (25)$$

guarantees that OLS selects a true column in the next iteration. Therefore, $\rho(\mathbf{r}_i) < 1$ for $i \in \{0, \dots, k-1\}$ ensures recovery of \mathbf{x} in k iterations. Equivalently, $\max_i \rho(\mathbf{r}_i) < 1$ is a sufficient condition. Therefore, if Σ denotes the event that OLS succeeds, then $\Pr\{\Sigma\} \geq \Pr\{\max_i \rho(\mathbf{r}_i) < 1\}$. One may upper bound $\rho(\mathbf{r}_i)$ as

$$\rho(\mathbf{r}_i) \leq \frac{\|\tilde{\mathbf{A}}^\top \mathbf{r}_i\|_\infty \max_{j \in \mathcal{S}_{\text{opt}}} \|\mathbf{P}_i^\perp \mathbf{a}_j\|_2}{\|\tilde{\mathbf{A}}^\top \mathbf{r}_i\|_\infty \min_{j \notin \mathcal{S}_{\text{opt}}} \|\mathbf{P}_i^\perp \mathbf{a}_j\|_2}. \quad (26)$$

According to Lemma III.2 and Lemma III.3,

$$\begin{aligned} \rho(\mathbf{r}_i) &\leq \frac{\|\tilde{\mathbf{A}}^\top \mathbf{r}_i\|_\infty \sqrt{1+\epsilon}}{\|\tilde{\mathbf{A}}^\top \mathbf{r}_i\|_\infty \sqrt{1-\epsilon}} \sqrt{\frac{(n-i)/n \mathbb{E} \|\mathbf{a}_{j_{\max}}\|_2}{(n-i)/n \mathbb{E} \|\mathbf{a}_{j_{\min}}\|_2}} \\ &= \sqrt{\frac{1+\epsilon}{1-\epsilon}} \frac{\|\tilde{\mathbf{A}}^\top \mathbf{r}_i\|_\infty}{\|\tilde{\mathbf{A}}^\top \mathbf{r}_i\|_\infty} \end{aligned} \quad (27)$$

with probability exceeding $p_1 = (1 - 2e^{-(n-k+1)c_0(\epsilon)})^2$ for $0 \leq i < k$. Let $c_1(\epsilon) = \sqrt{\frac{1-\epsilon}{1+\epsilon}}$. Following the framework in [28], using a simple norm inequality and exploiting the fact that $\tilde{\mathbf{A}}^\top \mathbf{r}_i$ has at most $k-i$ nonzero entries leads to

$$\rho(\mathbf{r}_i) \leq \frac{\sqrt{k-i}}{c_1(\epsilon)} \frac{\|\tilde{\mathbf{A}}^\top \mathbf{r}_i\|_\infty}{\|\tilde{\mathbf{A}}^\top \mathbf{r}_i\|_2} = \frac{\sqrt{k-i}}{c_1(\epsilon)} \|\tilde{\mathbf{A}}^\top \tilde{\mathbf{r}}_i\|_\infty, \quad (28)$$

where $\tilde{\mathbf{r}}_i = \mathbf{r}_i / \|\tilde{\mathbf{A}}^\top \mathbf{r}_i\|_2$. It is shown in [42] that for any $0 < \delta < 1$, $\Pr\{\|\tilde{\mathbf{r}}_i\|_2 \leq \frac{1}{1-\delta}\} \geq 1 - 2(\frac{12}{\delta})^k e^{-nc_0(\frac{\delta}{2})} = p_2$.

Subsequently,

$$\begin{aligned}
\Pr\{\Sigma\} &\geq p_1 p_2 \Pr\left\{\max_{0 \leq i < k} \|\tilde{\mathbf{A}}^\top \tilde{\mathbf{r}}_i \sqrt{k-i}\|_\infty < c_1(\epsilon)\right\} \\
&\geq p_1 p_2 \prod_{j=k+1}^m \Pr\left\{\max_{0 \leq i < k} \left|\mathbf{a}_j^\top \tilde{\mathbf{r}}_i \sqrt{k-i}\right| < c_1(\epsilon)\right\} \\
&= p_1 p_2 \Pr\left\{\max_{0 \leq i < k} \left|\mathbf{a}_j^\top \tilde{\mathbf{r}}_i \sqrt{k-i}\right| < c_1(\epsilon)\right\}^{m-k}
\end{aligned} \tag{29}$$

where we used the assumption that the columns of $\tilde{\mathbf{A}}$ are independent. Note that the random vectors $\{\tilde{\mathbf{r}}_i \sqrt{k-i}\}_{i=0}^{k-1}$ are bounded with probability exceeding p_2 and are statistically independent of $\tilde{\mathbf{A}}$. Now consider the case $\mathbf{A} \sim \mathcal{N}(0, \frac{1}{n})$. Since the random variable $X_{i,j} = \mathbf{a}_j^\top \tilde{\mathbf{r}}_i \sqrt{k-i}$ is distributed as $\mathcal{N}(0, \frac{k-i}{n(1-\delta)^2})$, by using a Gaussian tail bound and Boole's inequality it is straightforward to show that

$$\Pr\left\{\max_{0 \leq i < k} |X_{i,j}| < c_1(\epsilon)\right\} \geq 1 - \sum_{i=0}^{k-1} e^{-\frac{n}{k-i} c_1(\epsilon)^2 (1-\delta)^2}. \tag{30}$$

For the second case where $\mathbf{A} \sim \mathcal{B}(\frac{1}{2}, \pm \frac{1}{\sqrt{n}})$, this probability decreases to $1 - 2 \sum_{i=0}^{k-1} e^{-\frac{n}{k-i} c_1(\epsilon)^2 (1-\delta)^2}$ owing to Hoeffding inequality. Thus, in both cases $\Pr\{\Sigma\} \geq p_1 p_2 p_3$, where $p_3 = \left(1 - 2 \sum_{i=0}^{k-1} e^{-\frac{n}{k-i} c_1(\epsilon)^2 (1-\delta)^2}\right)^{m-k}$. This completes the proof.

APPENDIX C

PROOF OF THEOREM III.6

Following along the steps in the proof of Theorem III.5, due to noise, $\tilde{\mathbf{A}}^\top \mathbf{r}_i$ in 27 now has at most k nonzero entries. Hence, after a simple modification in 28, we obtain

$$\rho(\mathbf{r}_i) \leq \frac{\sqrt{k}}{c_1(\epsilon)} \|\tilde{\mathbf{A}}^\top \tilde{\mathbf{r}}_i\|_\infty. \tag{31}$$

Nevertheless, the most important difference is that \mathbf{r}_i in the noisy scenario is not in the range of $\tilde{\mathbf{A}}$ anymore and one can not simply employ results of [42]. Therefore, further precautions are needed to ensure that $\{\tilde{\mathbf{r}}_i\}_{i=0}^{k-1}$ remain bounded. Consequently, first we explore a lower bound for $\|\tilde{\mathbf{A}}^\top \mathbf{r}_i\|_2$ and then we bound $\tilde{\mathbf{r}}_i$ from above.

Recall that for i^{th} iteration,

$$\mathbf{r}_i = \mathbf{P}_i^\perp \mathbf{y} = \mathbf{P}_i^\perp (\tilde{\mathbf{A}} \bar{\mathbf{x}} + \boldsymbol{\nu}) \tag{32}$$

where $\bar{\mathbf{x}}$ corresponds to nonzero components of \mathbf{x} . Write $\boldsymbol{\nu}$ equivalently as

$$\boldsymbol{\nu} = \tilde{\mathbf{A}} \mathbf{w} + \boldsymbol{\nu}^\perp \tag{33}$$

where $\boldsymbol{\nu}^\perp = \mathbf{P}_k^\perp \boldsymbol{\nu}$ is projection of $\boldsymbol{\nu}$ onto orthogonal complement of subspace spanned by true columns, and $\mathbf{w} = \tilde{\mathbf{A}}^\dagger \boldsymbol{\nu}$. Substituting 33 into 32, noting $\mathbf{P}_i^\perp \mathbf{a} = 0$ if \mathbf{a} is selected in previous iterations, and the fact that $\mathcal{L}_i \subset \mathcal{L}_k$ we obtain

$$\mathbf{r}_i = \boldsymbol{\nu}^\perp + \mathbf{P}_i^\perp \tilde{\mathbf{A}}_{i^c} \mathbf{c}_{i^c} \tag{34}$$

where $\mathbf{c} = \bar{\mathbf{x}} + \mathbf{w}$, and subscript i^c denotes the set of optimal columns that have not been chosen yet. Evidently, 34 demonstrates \mathbf{r}_i as sum of orthogonal terms. Therefore,

$$\|\mathbf{r}_i\|_2^2 = \|\boldsymbol{\nu}^\perp\|_2^2 + \|\mathbf{P}_i^\perp \tilde{\mathbf{A}}_{i^c} \mathbf{c}_{i^c}\|_2^2. \tag{35}$$

Applying 34 in norm of $\tilde{\mathbf{A}} \mathbf{r}_i$ delivers,

$$\begin{aligned}
\|\tilde{\mathbf{A}} \mathbf{r}_i\|_2 &= \|\tilde{\mathbf{A}} (\boldsymbol{\nu}^\perp + \mathbf{P}_i^\perp \tilde{\mathbf{A}}_{i^c} \mathbf{c}_{i^c})\|_2 \\
&\stackrel{(a)}{=} \|\tilde{\mathbf{A}} \boldsymbol{\nu}^\perp + \tilde{\mathbf{A}}_{i^c} \mathbf{P}_i^\perp \tilde{\mathbf{A}}_{i^c} \mathbf{c}_{i^c}\|_2 \\
&\stackrel{(b)}{=} \|\tilde{\mathbf{A}}_{i^c} \mathbf{P}_i^\perp \tilde{\mathbf{A}}_{i^c} \mathbf{c}_{i^c}\|_2 \\
&\stackrel{(c)}{\geq} \sigma_{\min}^2(\tilde{\mathbf{A}}) \|\mathbf{c}_{i^c}\|_2
\end{aligned} \tag{36}$$

where (a) holds because \mathbf{P}_i^\perp projects onto orthogonal complement of column span of $\tilde{\mathbf{A}}_i$, (b) follows from the fact that columns of $\tilde{\mathbf{A}}$ and $\boldsymbol{\nu}^\perp$ lie in orthogonal subspaces, and (c) is from lemma III.4 and the fact that \mathbf{P}_i^\perp is a projection matrix.

We now bound norm of $\tilde{\mathbf{r}}_i$. Substitute 35 and 36 in definition of $\tilde{\mathbf{r}}_i$ to reach

$$\begin{aligned}
\|\tilde{\mathbf{r}}_i\|_2 &\leq \frac{[\|\boldsymbol{\nu}^\perp\|_2^2 + \|\mathbf{P}_i^\perp \tilde{\mathbf{A}}_{i^c} \mathbf{c}_{i^c}\|_2^2]^{\frac{1}{2}}}{\sigma_{\min}^2(\tilde{\mathbf{A}}) \|\mathbf{c}_{i^c}\|_2} \\
&\stackrel{(a)}{\leq} \frac{[\|\boldsymbol{\nu}^\perp\|_2^2 + \sigma_{\max}^2(\tilde{\mathbf{A}}) \|\mathbf{c}_{i^c}\|_2^2]^{\frac{1}{2}}}{\sigma_{\min}^2(\tilde{\mathbf{A}}) \|\mathbf{c}_{i^c}\|_2} \\
&= \frac{[\|\boldsymbol{\nu}^\perp\|_2^2 / \|\mathbf{c}_{i^c}\|_2^2 + \sigma_{\max}^2(\tilde{\mathbf{A}})]^{\frac{1}{2}}}{\sigma_{\min}^2(\tilde{\mathbf{A}})}
\end{aligned} \tag{37}$$

where (a) is from lemma III.4 and the fact that \mathbf{P}_i^\perp is a projection matrix. In addition,

$$\|\boldsymbol{\nu}^\perp\|_2 = \|\mathbf{P}_k^\perp \boldsymbol{\nu}\|_2 \leq \|\boldsymbol{\nu}\|_2 \leq \epsilon_\nu \tag{38}$$

On the other hand, with defining $\mathbf{x}_{\min} = \min_j |\bar{\mathbf{x}}_j|$ and $\mathbf{c}_{\min} = \min_j |\mathbf{c}_j|$ it is easy to check

$$\mathbf{c}_{\min} \geq \mathbf{x}_{\min} - \|\mathbf{w}\|_2. \tag{39}$$

We further impose $\mathbf{x}_{\min} \geq (1 + \delta) \|\mathbf{w}\|_2$. Therefore,

$$\begin{aligned}
\|\mathbf{c}_{i^c}\|_2^2 &\geq (k-i) \mathbf{c}_{\min}^2 \\
&\geq (k-i) (\mathbf{x}_{\min} - \|\mathbf{w}\|_2)^2 \\
&= (k-i) (\mathbf{x}_{\min} - \|\tilde{\mathbf{A}}^\dagger \boldsymbol{\nu}\|_2)^2 \\
&\geq (k-i) (\mathbf{x}_{\min} - \sigma_{\max}(\tilde{\mathbf{A}}^\dagger) \|\boldsymbol{\nu}\|_2)^2 \\
&= (k-i) (\mathbf{x}_{\min} - \sigma_{\min}(\tilde{\mathbf{A}}) \epsilon_\nu)^2.
\end{aligned} \tag{40}$$

Combining 37, 38, and 40 furnishes

$$\begin{aligned}
\|\tilde{\mathbf{r}}_i\|_2 &\leq \frac{\left[\frac{\epsilon_\nu^2}{(k-i)(\mathbf{x}_{\min} - \sigma_{\min}(\tilde{\mathbf{A}}) \epsilon_\nu)^2} + \sigma_{\max}^2(\tilde{\mathbf{A}}) \right]^{\frac{1}{2}}}{\sigma_{\min}^2(\tilde{\mathbf{A}})} \\
&\leq \frac{\left[\frac{\epsilon_\nu^2}{(k-i)(\mathbf{x}_{\min} - (1+\delta)\epsilon_\nu)^2} + (1+\delta)^2 \right]^{\frac{1}{2}}}{(1-\delta)^2}
\end{aligned} \tag{41}$$

with probability exceeding p_2 . Thus, imposing the constraint

$$\mathbf{x}_{\min} \geq (1 + \delta + t) \epsilon_\nu \tag{42}$$

for any $t > 0$ ⁵ establishes

$$\|\tilde{\mathbf{r}}_i\|_2 \leq \frac{\left[\frac{1}{(k-i)t^2} + (1+\delta)^2 \right]^{\frac{1}{2}}}{(1-\delta)^2}. \tag{43}$$

⁵Note that this agrees with our restriction of $\mathbf{x}_{\min} \geq (1 + \delta) \|\mathbf{w}\|_2$.

Following the steps in proof of theorem III.5 and using the independence assumption of columns of $\tilde{\mathbf{A}}$,

$$\Pr\{\Sigma\} \geq p_1 p_2 \Pr\left\{\max_{0 \leq i < k} |\mathbf{a}_j^\top \tilde{\mathbf{r}}_i| < \frac{c_1(\epsilon)}{\sqrt{k}}\right\}^{m-k}. \quad (44)$$

Given that $\{\tilde{\mathbf{r}}_i\}_{i=0}^{k-1}$ are bounded with probability higher than p_2 and are statistically independent of $\tilde{\mathbf{A}}$, by applying Boole's inequality or Hoeffding inequality, depending on whether $\mathbf{A} \sim \mathcal{N}(0, \frac{1}{n})$ or $\mathbf{A} \sim \mathcal{B}(\frac{1}{2}, \pm \frac{1}{\sqrt{n}})$, for random variable $X_{i,j} = \mathbf{a}_j^\top \tilde{\mathbf{r}}_i$ we obtain,

$$\Pr\left\{\max_{0 \leq i < k} |X_{i,j}| < \frac{c_1(\epsilon)}{\sqrt{k}}\right\} \geq 1 - 2 \sum_{i=0}^{k-1} e^{-\frac{nc_1(\epsilon)^2(1-\delta)^4}{k \left[\frac{1}{(k-i)^2} + (1+\delta)^2 \right]}}. \quad (45)$$

Denote

$$p_3 = \left(1 - 2 \sum_{i=0}^{k-1} e^{-\frac{nc_1(\epsilon)^2(1-\delta)^4}{k \left[\frac{1}{(k-i)^2} + (1+\delta)^2 \right]}}\right)^{m-k} \quad (46)$$

From 44, and 45 follows $\Pr\{\Sigma\} \geq p_1 p_2 p_3$ which completes the proof.

Remark 4: It is worth comparing 43 with that of noiseless scenario. Without noise, the first term in the numerator of 43 vanishes and we obtain $\|\tilde{\mathbf{r}}_i\|_2 \leq \frac{1}{1-\delta} + \frac{2\delta}{(1-\delta)^2}$ where the term $\frac{2\delta}{(1-\delta)^2}$ is the expense we pay as a result of taking noise into account in our estimates.

APPENDIX D PROOF OF THEOREM IV.1

Assume column \mathbf{a}_{j_s} is selected in $(i+1)^{\text{st}}$ iteration. We begin by post-multiplying both sides of 5 with observation vector \mathbf{y} :

$$\mathbf{P}_{i+1}^\perp \mathbf{y} = \mathbf{P}_i^\perp \mathbf{y} - \frac{\mathbf{P}_i^\perp \mathbf{a}_{j_s} \mathbf{a}_{j_s}^\top \mathbf{P}_i^\perp}{\|\mathbf{P}_i^\perp \mathbf{a}_{j_s}\|_2^2} \mathbf{y}. \quad (47)$$

By definition of the residual vector, i.e., $\mathbf{r}_i = \mathbf{P}_i^\perp \mathbf{y}$ and definition of z_{j_s} in 18 we obtain

$$\begin{aligned} \mathbf{r}_{i+1} &= \mathbf{r}_i - \frac{\mathbf{P}_i^\perp \mathbf{a}_{j_s}}{\|\mathbf{P}_i^\perp \mathbf{a}_{j_s}\|_2^2} \mathbf{a}_{j_s}^\top \mathbf{r}_i \\ &= \mathbf{r}_i - \frac{\mathbf{P}_i^\perp \mathbf{a}_{j_s}}{\|\mathbf{P}_i^\perp \mathbf{a}_{j_s}\|_2^2} z_{j_s}. \end{aligned} \quad (48)$$

In order to complete the proof, one needs to show $\mathbf{u}_{i+1} = \frac{\mathbf{P}_i^\perp \mathbf{a}_{j_s}}{\|\mathbf{P}_i^\perp \mathbf{a}_{j_s}\|_2^2}$. Owing to the fact that OLS does not identify repeated columns and the assumption that \mathbf{A} is full rank, it is evident that previously selected columns are linearly independent. Specifically, let $\{\tilde{\mathbf{a}}_l\}_{l=1}^i$ be the sequence of selected columns in the first i iterations and that $\mathcal{L}_i = \{\tilde{\mathbf{a}}_1, \dots, \tilde{\mathbf{a}}_i\}$ is the subspace spanned by these columns. Now, consider the term $\mathbf{P}_i^\perp \mathbf{a}_{j_s}$ which is the orthogonal projection of the selected column \mathbf{a}_{j_s} onto \mathcal{L}_i . It is easy to see that $\mathbf{P}_i^\perp \mathbf{a}_{j_s} = \mathbf{a}_{j_s} - \mathbf{P}_i \mathbf{a}_{j_s}$. Comparing with definition of \mathbf{t} in 18, it suffices to demonstrate $\mathbf{P}_i \mathbf{a}_{j_s} = \sum_{l=1}^i \frac{\mathbf{a}_{j_s}^\top \mathbf{u}_l}{\|\mathbf{u}_l\|_2^2} \mathbf{u}_l$. Since $\{\tilde{\mathbf{a}}_l\}_{l=1}^i$ are linearly independent, one can easily construct an orthogonal basis for \mathcal{L}_i using a procedure similar to the so-called Modified Gram-Schmidt (MGS) method. Because of the

structure of 5, i.e., squared l_2 -norm of $\mathbf{P}_i^\perp \mathbf{a}_{j_s}$ in the denominator, orthogonalized columns are divided by their squared l_2 -norm rather than being normalized. For instance, the very first vector in the orthogonal basis would be $\mathbf{u}_1 = \frac{\tilde{\mathbf{a}}_1}{\|\tilde{\mathbf{a}}_1\|_2}$. Consequently, orthogonal projection of \mathbf{a}_{j_s} is acquired by Euclidean projection of \mathbf{a}_{j_s} onto each of the orthogonal vectors \mathbf{u}_l . Therefore, $\mathbf{P}_i \mathbf{a}_{j_s} = \sum_{l=1}^i \frac{\mathbf{a}_{j_s}^\top \mathbf{u}_l}{\|\mathbf{u}_l\|_2^2} \mathbf{u}_l$ holds which completes the proof.

REFERENCES

- [1] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *The journal of machine learning research*, vol. 1, pp. 211–244, 2001.
- [2] D. L. Donoho, "Compressed sensing," *Information Theory, IEEE Transactions on*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [3] C. Carbonelli, S. Vedantam, and U. Mitra, "Sparse channel estimation with zero tap detection," *IEEE Transactions on Wireless Communications*, vol. 6, no. 5, pp. 1743–1763, 2007.
- [4] S. Barik and H. Vikalo, "Sparsity-aware sphere decoding: algorithms and complexity analysis," *Signal Processing, IEEE Transactions on*, vol. 62, no. 9, pp. 2212–2225, 2014.
- [5] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, "Recovering sparse signals using sparse measurement matrices in compressed dna microarrays," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 2, no. 3, pp. 275–285, 2008.
- [6] R. G. Baraniuk, "Compressive sensing," *IEEE signal processing magazine*, vol. 24, no. 4, 2007.
- [7] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging," *Magnetic resonance in medicine*, vol. 58, no. 6, pp. 1182–1195, 2007.
- [8] M. Elad, M. A. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, 2010.
- [9] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-nyquist sampling of sparse wideband analog signals," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 2, pp. 375–391, 2010.
- [10] Y. C. Eldar and G. Kutyniok, *Compressed sensing: theory and applications*. Cambridge University Press, 2012.
- [11] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2790–2797.
- [12] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.
- [13] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.
- [14] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.
- [15] Y. C. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, 1993, pp. 40–44.
- [16] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 58, no. 2, pp. 1094–1121, 2012.
- [17] J. Wang, S. Kwon, and B. Shim, "Generalized orthogonal matching pursuit," *Signal Processing, IEEE Transactions on*, vol. 60, no. 12, pp. 6202–6216, 2012.
- [18] D. Needell and J. A. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.
- [19] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *Information Theory, IEEE Transactions on*, vol. 55, no. 5, pp. 2230–2249, 2009.
- [20] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations of computational mathematics*, vol. 9, no. 3, pp. 317–334, 2009.

- [21] T. Zhang, "Sparse recovery with orthogonal matching pursuit under rip," *Information Theory, IEEE Transactions on*, vol. 57, no. 9, pp. 6215–6221, 2011.
- [22] M. A. Davenport and M. B. Wakin, "Analysis of orthogonal matching pursuit using the restricted isometry property," *Information Theory, IEEE Transactions on*, vol. 56, no. 9, pp. 4395–4401, 2010.
- [23] Q. Mo and Y. Shen, "A remark on the restricted isometry property in orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 58, no. 6, pp. 3654–3656, 2012.
- [24] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *Information Theory, IEEE Transactions on*, vol. 50, no. 10, pp. 2231–2242, 2004.
- [25] T. T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *Information Theory, IEEE Transactions on*, vol. 57, no. 7, pp. 4680–4688, 2011.
- [26] T. Zhang, "On the consistency of feature selection using greedy least squares regression," in *Journal of Machine Learning Research*, 2009, pp. 555–568.
- [27] A. Das and D. Kempe, "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1057–1064.
- [28] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *Information Theory, IEEE Transactions on*, vol. 53, no. 12, pp. 4655–4666, 2007.
- [29] S. Rangan and A. K. Fletcher, "Orthogonal matching pursuit from noisy random measurements: A new analysis," in *Advances in Neural Information Processing Systems*, 2009, pp. 540–548.
- [30] A. K. Fletcher and S. Rangan, "Orthogonal matching pursuit: A brownian motion analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1010–1021, 2012.
- [31] N. R. Draper, H. Smith, and E. Pownell, *Applied regression analysis*. Wiley New York, 1966, vol. 3.
- [32] R. Hocking, "The analysis and selection of variables in linear regression. biometrics 32 431–453," *Mathematical Reviews (MathSciNet): MR398008 Digital Object Identifier: doi*, vol. 10, p. 2529336, 1976.
- [33] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of control*, vol. 50, no. 5, pp. 1873–1896, 1989.
- [34] C. Soussen, R. Gribonval, J. Idier, and C. Herzet, "Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares," *Information Theory, IEEE Transactions on*, vol. 59, no. 5, pp. 3158–3174, 2013.
- [35] C. Herzet, A. Drémeau, and C. Soussen, "Relaxed recovery conditions for omp/ols by exploiting both coherence and decay," *Information Theory, IEEE Transactions on*, vol. 62, no. 1, pp. 459–470, 2016.
- [36] C. Herzet, C. Soussen, J. Idier, and R. Gribonval, "Exact recovery conditions for sparse representations with partial support information," *Information Theory, IEEE Transactions on*, vol. 59, no. 11, pp. 7509–7524, 2013.
- [37] A. Hashemi and H. Vikalo, "Sparse linear regression via generalized orthogonal least-squares," *arXiv preprint arXiv:1602.06916*, 2016.
- [38] D. Achlioptas, "Database-friendly random projections," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 2001, pp. 274–281.
- [39] S. Dasgupta and A. Gupta, "An elementary proof of the johnson-lindenstrauss lemma," *International Computer Science Institute, Technical Report*, pp. 99–006, 1999.
- [40] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [41] M. Grant and S. Boyd, "Cvx: Matlab software for disciplined convex programming."
- [42] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.